Amendments to the Specification:

Please add the following new paragraphs beginning at page 35, line 28 of the Substitute Specification filed on June 7, 2002:

Section I: expanded methods

Patient data and tumor bank

The complete cohort for these studies consists of 68 children with medulloblastomas, 10 young adults with malignant gliomas (WHO grades III and IV), 5 children with AT/RT, 5 with renal/extrarenal rhabdoid tumors, and 8 children with supratentorial PNETs. A summary of the clinical data for the patients can be found in the List of all samples section of the document. All patients with medulloblastomas were treated with craniospinal irradiation to 2400 - 3600 centiGray (cGy) with a tumor dose of 5300 - 7200 cGy. All patients with medulloblastomas were treated with chemotherapy consisting of cisplatin and vincristine, and combinations of carboplatin, etoposide, cyclophosphamide, procarbor lomustine (CCNU). Two patients received high dose chemotherapy at relapse, including methotrexate and thiotepa, followed by autologous bone marrow transplantation. Thirty-five of the children with medulloblastomas were part of a cohort described in previous publications (Segal *et al.*, 1994, Kim *et al.*, 1999). All tumor samples were obtained at the time of initial surgery prior to treatment. The samples were snap frozen in liquid nitrogen and stored at -80°C. The studies were done with approval of the Committee for Clinical Investigation of Boston Children's Hospital. The data were organized into three sets: Dataset A (42 samples containing: 10 medulloblastomas, 10 malignant gliomas, 5 AT/RT and 5 renal/extrarenal rhabdoid tumors, 8 supratentorial PNETs and 4 normal cerebella), Dataset B (34 samples, containing 9 desmoplastic medulloblastoma and 25 classic medulloblastoma), and Dataset C (60 samples, containing 39 medulloblastoma survivors and 21 treatment failures). There are two additional variants of Dataset A called A1 and A2. A description of each dataset is available in the Datasets and clinical attributes.

Microarray hybridization

Tissue samples were homogenized (Polytron, Kinematica, Lucerne) in guanidinium isothiocyanate and RNA was isolated by centrifugation over a CsCl gradient. RNA integrity was assessed either by northern blotting (Kim *et al.*, 1999) or by gel electrophoresis. The amount of

starting total RNA for each reaction varied between 10 and 12 μg. First strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An *in vitro* transcription reaction was done to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C for 35 minutes. Ten micrograms of the fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-Morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin (Sigma, St. Louis) to Affymetrix (Santa Clara, CA) HuGeneFL arrays at 45°C for 16 hours

HuGeneFL arrays contain 5920 known genes and 897 expressed sequence tags. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) at 3 μg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent. Scans were performed on Affymetrix scanners and the expression value for each gene was calculated using Affymetrix GENECHIP software. Minor differences in microarray intensity were corrected using a linear scaling method as detailed in the next section.

Preprocessing and re-scaling

The raw expression data as obtained from Affymetrix's GeneChip is re-scaled to account for different chip intensities. Each column (sample) in the dataset was multiplied by 1/slope of a least squares linear fit of the sample vs. the reference (the first sample in the dataset). This linear fit is done using only genes that have 'Present' calls in both the sample being re-scaled and the reference. The sample chosen as reference is a typical one (*i.e.*, one with the number of "P" calls closer to the average over all samples in the dataset). Scans were rejected if the scaling factor exceeded a factor of 3, fewer than 1000 genes received 'Present' calls, or microarray artifacts were visible.

A ceiling of 16,000 units was chosen for all experiments because it is at this level that we observed fluorescence saturation of the scanner; values above this cannot be reliably measured. For classification problems that are very robust (*e.g.*, distinguishing different types of brain tumors), we used a threshold of 100 units because there was a sufficiently large number of genes correlated with the distinction that the threshold could be set high, thereby minimizing noise, and maximizing potential biological interpretation of the marker genes. For the more subtle

distinctions (*e.g.*, outcome prediction), few correlates of the distinction are found, and for this reason the threshold was set at a lower level (20 units) so as to avoid missing any potentially informative marker genes.

These numbers are Affymetrix's scanner "average difference" units. After this preprocessing gene expression values were subjected to a variation filter which excluded genes showing minimal variation across the samples being analyzed. The variation filter tests for a fold-change and absolute variation over samples (comparing max/min and max-min with predefined values and excluding genes not obeying both conditions). The precise parameters of the variation filters for each dataset are provided in each analysis section of this document. Different thresholds and variation filters were used according to the purpose of the analysis (*e.g.*, select weak marker genes for treatment outcome, strong robust marker genes for morphology, highly varying genes for PCA etc.). For example, if the maximum and minimum values of a gene across samples were max and min then the variation filter excluded those where max/min < 5 and max - min < 500. In some cases more or less stringent values were used.

Clustering

Self Organizing Maps were performed using our GeneCluster clustering package.

Self-Organizing Maps (SOMs). The Self Organizing Map is a method for performing unsupervised learning (*i.e.*, learning models for classifying data where the true class for the data samples is assumed to be unknown prior to model training) where a grid of 2D nodes (clusters) is iteratively adjusted to reflect the global structure in the expression dataset (Tamayo *et al.*, 1999). In general, unsupervised learning presents a more difficult problem than supervised learning methods (such as weighted voting or k-NN) but is useful for discovering new classes during exploratory analysis. With the SOM, one randomly chooses the geometry of the grid (*e.g.*, a 3 x 2 grid) and maps it into the k-dimensional feature space. Initially the features are randomly mapped to the grid but during training the mapping is iteratively adjusted to reflect the data structure. The data were first normalized by standardizing each column (sample) to mean 0 and variance 1. The SOM results for the clustering of samples can be found in the Multiple tumor clustering for multiple tumor samples and in the SOM clustering of treatment outcome samples.

Hierarchical Clustering is another unsupervised learning method useful for dividing data into natural groups. Data is clustered hierarchically by organizing the data into a tree structure based upon the degree of similarity between features. We used the Cluster and TreeView software (Eisen *et al.*, 1998) to perform average linkage clustering, which organizes all of the data elements into a single tree with the highest levels of the tree representing the discovered classes. The detailed clustering results can be accessed in the Multiple tumor clustering section.

Description of the permutation test-based neighborhood analysis method

Permutation test based (Golub *et al.*, 1999) neighborhood analysis is used to select and screen marker genes with respect to biologically meaningful phenotypes (morphology and treatment outcome) and to assess their statistical significance. To accomplish this we compare the top signal-to-noise scores of top marker genes with the corresponding ones from data obtained by randomly permuting the class labels. Typically 500 global random permutations were used to build histograms. Based on these histograms we determined the 50% (median), 5% and 1% significance levels and compared them with the values obtained for the real dataset. As described above this procedure is motivated by considering the following question: what is the likelihood that a given set of markers genes, for example selected by signal to noise, of a phenotype of interest represent chance correlations and not biologically significant matches? If one looks down the list of markers, how many should one consider as input to a classifier or for further study? In this list of selected markers what is the best way to minimize the number of false positives but retain enough sensitivity to select a non-empty set?

In detail the permutation test procedure for a given comparison of interest (*e.g.*, markers high in class 0 and low in class 1) is as follows:

Generate signal-to-noise $(\mu_{class\ 0} - \mu_{class\ 1})/(\sigma_{class\ 0} + \sigma_{class\ 1})$ scores for all genes that pass a variation filter using the actual class labels (phenotype) and sort them accordingly. The best match (k = 1) is the gene "closer" or more correlated to the phenotype using the signal to noise as a correlation function. In fact one can imagine the reciprocal of the signal to noise as a "distance" between the "phenotype" and each gene. One can also use a *t*-statistic $(\mu_{class\ 0} - \mu_{class\ 1})/(\sigma_{class\ 0} + \sigma_{class\ 1})^{1/2}$ and obtain very similar results.

Generate 500 or more random permutations of the class labels (phenotype). For each case of randomized class labels generate signal-to-noise scores and sort genes accordingly.

Build a histogram of signal to noise scores for each value of k. For example one for all the 500 top markers (k = 1), another one for the 500 second best (k = 2), etc. These histograms represent a reference statistic for the best match, second best, etc. and, for a given value of k, different genes contribute to it. Notice that the correlation structure of the data is preserved by this procedure. For each value of k, determine different percentiles (1%, 5%, 50% etc.) of the corresponding histogram.

Compare the actual signal to noise scores with the different significance levels obtained for the histograms of permuted class labels for each value of k. This test helps to assess the statistical significance of gene markers in terms of the distribution of class-gene scores using permuted labels.

Algorithms

k-Nearest Neighbors (k-NN)

We developed a weighted implementation of the k-NN algorithm (Dasarathy, 1991) that predicts the class of a new sample by calculating the Euclidean distance (d) of this sample to the k "nearest neighbor" standardized samples in "expression" space in the training set, and by selecting the predicted class to be that of the majority of the k samples (the method is defined in terms of Euclidean distances over standardized vectors so it is equivalent to using inner products: $a \cdot b / |a||b|$). We performed the marker gene selection process by which we feed the k-NN algorithm only the features with higher correlation with the target class. This feature selection is done by sorting the features according to the signal-to-noise statistic (Golub 1999, Slonim 2000) $(\mu_{class\ 0} - \mu_{class\ 1})/(\sigma_{class\ 0} + \sigma_{class\ 1})$. In our version of the algorithm the weight of each of the k neighbors was weighted according to 1/d. For our medulloblastoma outcome experiments, the k-NN models were evaluated by 60-fold leave-one-out cross-validation whereby a training set of 59 samples was used to predict the class of a randomly withheld sample. This was repeated for

all samples and the cumulative error rate was recorded. Models with variable numbers of genes (1-200, selected according to their correlation with the survivor vs. treatment failure distinction in the training set) were tested in this manner.

Weighted Voting

The weighted voting algorithm (Golub 1999, Slonim 2000) makes a weighted linear combination of relevant "marker" or "informative" genes obtained in the training set to provide a classification scheme for new samples. The selection of features (marker genes) is accomplished by computing the signal-to-noise statistic S_x (described above). The class predictor is uniquely defined by the initial set of samples and marker genes. In addition to computing S_x, the algorithm also finds the decision boundaries (half way) between the class means: $b_x = (\mu_{class\ 0} + \mu_{class\ 1})/2$ for each gene. To predict the class of a test sample y, each gene x in the feature set casts a vote: $V_x = S_x\ (g_x^y - b_x)$ and the final vote for class 0 or 1 is sign $(\Sigma_x\ V_x)$. The strength or confidence in the prediction of the winning class is $(V_{win} - V_{lose})/(V_{win} + V_{lose})$ (*i.e.*, the relative margin of victory for the vote). The detailed prediction results are the Weighted voting treatment outcome prediction results.

Support Vector Machines

The Support Vector Machine (SVM) for classification minimizes the generalization error rather than the training error. The basic idea behind SVMs is to construct an optimal separating hyperplane by mapping the gene expression data to a high-dimensional space (Mukherjee *et al.*, 1999, Brown *et al.*, 2000). Linear separation in this higher dimensional space corresponds to a nonlinear decision boundary in the original space. A new feature selection algorithm was developed to scale the input features to minimize the ratio of the radius around the support vectors and the margin.

SPLASH

The Splash algorithm (Califano *et al.*, 1999) discovers efficiently and deterministically all statistically significant gene expression patterns in a target class of interest. Statistical significance is evaluated based on the probability of a "pattern," (*i.e.*, a subset of genes and

experiments within a narrow interval of expression values) to occur by chance in the control target class. A greedy set covering algorithms is used to select an optimal subset of statistically significant patterns. These patterns are accumulated and form the basis for a likelihood ratio classification scheme to predict new samples. The detailed results are in the SPLASH treatment outcome prediction results section.

Predictors using metastatic staging and TrkC

These classifiers were constructed by finding the decision boundary half way between the classes: $(\mu_{class\ 0} + \mu_{class\ 1})/2$ (using the staging values 0 vs. 1,2,3,4 or the continuous TrkC gene expression) and then predicting the unknown sample according to its gene expression value location with respect to that boundary. The detailed results can be found in the TrkC treatment outcome prediction results and Staging treatment outcome prediction results sections.

Proportional chance criterion.

In order to compute p-values for non-survival predictions, for example the p-val = 4×10^{-7} for the Classic vs. Desmoplastic classifier reported in the paper (33 out of 34 samples correctly classified) we used a "proportional chance criterion" to evaluate the probability that a random predictor will produce a confusion matrix with the same row and column counts as the gene expression predictor. For example, for a binary class (A vs. B) problem, if a is the prior probability of a sample being in class A and p is the true proportion of samples in class A then $C_p = p\,\alpha + (1 - p)(1 - \alpha)$ is the proportion of the overall sample that is expected to receive correct classification by chance alone. Then if C_{model} is the proportion of correct classifications achieved by the gene expression predictor one can estimate its significance by using a Z statistic of the form: $(C_{model} - C_p)/\mathrm{Sqrt}(C_p\,(1 - C_p)/n)$, where n is the total sample count. For more details see chapter VII of Huberty 1994.

Survival analysis and Kaplan-Meier plots

The Kaplan-Meier survival analysis plots are computed using the S-Plus (at the website, insightful.com/products/splus/) statistical software package: S-Plus 2000, Guide to Statistics Volume 2, chapter 9. The p-values for the prediction of outcome groups are computed using a

log-rank test (Mantel-Haenszel method, chapter 9 in the same reference). The Kaplan Meier plots and associated rank test p-values are included at the end of each of the outcome prediction sections starting in the k-nearest neighbors treatment outcome prediction results section.

PCA and multidimensional-scaling of Brain tumor samples

Datasets of large dimensionality (*i.e.*, large number of variables, *e.g.*, genes) are in general difficult to visualize due to the intrinsic difficulty of reducing and projecting the dataset to a small number of dimensions where standard visualization techniques are applicable. The main problem of performing a projection of that sort is that of preserving the "relevant" or "interesting" structure in the data. In our case this structure corresponds to the intrinsic similarities or the natural clustering of brain samples in the space of gene expression.

A commonly used technique for data reduction, projection and visualization is Principal Component Analysis (PCA). In this approach one finds standardized linear combinations of variables, the "principal components," which are orthogonal and explain all of the variance in the original dataset. A typical method to obtain a simple projection (multi-dimensional scaling) of the dataset is to plot the top 2 or 3 principal components, which may account for a significant fraction of the variance, in a 2 or 3D scatter plot.

To study the natural clustering of the Brain tumor samples we performed PCA analysis and projected the top three components in 3D and 2D scatter plots. We considered two subsets of genes: highly varying, those with highest variation across samples that passed a variation filter (1,065 genes) and, marker genes, the top 10 marker genes of each tumor class by using the signal-to-noise statistics as described in the statistical analysis and prediction section. For the highest variation genes the values were thresholded to 100 from below and 16,000 from above and the variation filter selected genes with at least a 12-fold and 1,200 absolute units of variation between the minimum and maximum values across samples. This produced a subset of 1,065 highly varying genes. For the marker genes the values were thresholded to 20 from below and 16,000 from above and a variation filter selected genes with at least a 5-fold and 500 absolute units of variation between the minimum and maximum values across samples. The genes that passed this filter were ranked according to signal to noise (using medians) and the top 10 markers for each class were selected. This produced a total of 50 genes.

Once the appropriate subset of highly varying or maker genes was selected we computed the 3 principal components using the S-Plus statistical software package using default settings. These three components were then plotted in 3D scatter plots. The plots show the "natural" clustering of brain tumor samples in these two subspaces of gene expression. The components and plots can also be seen in the Multiple tumor PCA section. Besides the 2D and 3D plots of the top 3 components we also include bar graphs showing the relative importance of the top components and the loadings of the top 6 genes for each component.

Combined classifiers

The fact that sometimes the prediction algorithms make mistakes in different samples and that the class structure of the confusion matrices is different for each algorithm motivated us to combine some of them to see if the predictions can be improved in this way. We choose a simple scheme combining three algorithms according to majority. For example if the outputs of the three algorithms for a given sample are Survivor, failure, and Survivor, then the output of the combined predictor will be Survivor. The results for two types of model combinations: using a simple majority rule: Staging, k-NN and TrkC and SVM, k-NN and TrkC can be seen in the Combined treatment outcome predictors section

Dataset A, A1, A2 - multiple tumor samples

Dataset A: 10 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 4 normal cerebellums and 8 supratentorial PNETs.

Two of the supratentorial PNETs are pineoblastomas, which historically have been inconsistently included in the PNET category. The analysis was repeated excluding these 2 pineoblastomas.

Dataset A1: 10 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 4 normal cerebellums and 6 supratentorial PNETs.

To test whether inclusion of a larger number of medulloblastomas might lessen the distinctions noted in Dataset A, 50 more medulloblastoma samples were added and the PCA analysis repeated.

Dataset A2: 60 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 4 normal cerebellums and 6 supratentorial PNETs.

Section II: datasets and clinical attributes

The following sections of this document describe the samples, clinical attributes and datasets in detail.

List of all samples

Number	Sample name	Type	Subtype	Chang Stage	Sex	Age at diagnosis [years/ months]	Followup [Months]	Current status [Alive/Dead]	Chemotherapy
1	Brain_MD_1	Medulloblastoma	Classic	T4M1	M	8m	11	D	V,C,Cx,VP
2	Brain_MD_2	Medulloblastoma	Classic	T2M0	M	8yr10m	5	D	V,C,Cx,VP
3	Brain_MD_3	Medulloblastoma	Classic	T3M0	M	6yr	7	D	V,C,Cx
4	Brain_MD_4	Medulloblastoma	Classic	T3M3	M	5yr 3m	7	D	V,C,Cx,VP
5	Brain_MD_5	Medulloblastoma	Classic	M3	M	38yr 2m	7	D	V,C
6	Brain_MD_6	Medulloblastoma	Classic	T4M0	F	7m	9	D	V,C,Cx
7	Brain_MD_7	Medulloblastoma	Classic	T1M0	M	6yr 5m	14	D	V,C,Cx
8	Brain_MD_8	Medulloblastoma	Classic	T3bM1	M	6yr 1m	16	D	V,C,Cx
9	Brain_MD_9	Medulloblastoma	Classic	M0	M	8yr	18	D	V,C,Cx,VP
10	Brain_MD_10	Medulloblastoma	Classic	M0	M	3yr 10m	18	D	V,C,Cx
11	Brain_MD_11	Medulloblastoma	Classic	T2M1	M	8yr 2m	19	D	V,C,Cx,VP,Ca,T,M
12	Brain_MD_12	Medulloblastoma	Classic	M0	F	3yr 9m	25	D	V,C,Cx
13	Brain_MD_13	Medulloblastoma	Classic	T3M3	M	14yr 5m	26	D	V,C,Cx
14	Brain_MD_14	Medulloblastoma	Desmoplastic	M0	M	6yr 3m	33	D	V,C,CC
15	Brain_MD_15	Medulloblastoma	Desmoplastic	T2MO	F	11yr 7m	38	D	V,C,Cx,VP
16	Brain_MD_16	Medulloblastoma	Desmoplastic	T3M3	F	11yr 5m	39	D	V,C,VP
17	Brain_MD_17	Medulloblastoma	Classic	T3bM3	F	3yr 3m	39	D	V,C,Cx
18	Brain_MD_18	Medulloblastoma	Classic	T2M3	M	4yr 4m	42	D	V,C,Cx
19	Brain_MD_19	Medulloblastoma	Classic	M2	F	26yr 1m	65	D	V,C,Cx,VP
20	Brain_MD_20	Medulloblastoma	Classic	T3bM0	M	20yr 6m	92	D	V,C
21	Brain_MD_21	Medulloblastoma	Classic	T2M0	F	23yr 3m	102	D	V,C
22	Brain_MD_22	Medulloblastoma	Desmoplastic	M0	F	5yr 7m	24	A	V,C,CC
23	Brain_MD_23	Medulloblastoma	Desmoplastic	T4M0	M	1yr 4m	25	A	V,C,Cx
24	Brain_MD_24	Medulloblastoma	Classic	T3M0	M	10yr 10m	27	A	V,C,Cx
25	Brain_MD_25	Medulloblastoma	Classic	M0	F	5yr 4m	28	A	V,C,Cx,VP
26	Brain_MD_26	Medulloblastoma	Classic	T2M3	M	1yr	33	A	V,C,Cx,VP
27	Brain_MD_27	Medulloblastoma	Classic	M0	M	5yr 10m	34	A	V,C,Cx

28	Brain_MD_28	Medulloblastoma	Desmoplastic	T4M0	M	6yr 1m	35	A	V,C,Cx
29	Brain_MD_29	Medulloblastoma	Classic	T3M0	F	7yr 5m	35	A	V,C,Cx
30	Brain_MD_30	Medulloblastoma	Desmoplastic	T3M0	F	11yr 9m	36	A	V,C,Cx
31	Brain_MD_31	Medulloblastoma	Classic	M0	M	7yr 4m	39	A	V,C,Cx
32	Brain_MD_32	Medulloblastoma	Desmoplastic	T2M0	M	10yr 11m	39	A	V,C,Cx
33	Brain_MD_33	Medulloblastoma	Classic	T3bM0	M	12yr 9m	41	A	V,C,Cx
34	Brain_MD_34	Medulloblastoma	Classic	T3M1	M	8yr 2m	42	A	V,C,Cx
35	Brain_MD_35	Medulloblastoma	Desmoplastic	T3M0	F	2yr 3m	45	A	V,C,Cx
36	Brain_MD_36	Medulloblastoma	Classic	T3M0	M	5yr 6m	46	A	V,C,Cx
37	Brain_MD_37	Medulloblastoma	Classic	T3M0	F	12yr 7m	51	A	V,C,Cx
38	Brain_MD_38	Medulloblastoma	Desmoplastic	T3M1	F	7m	52	A	V,C,Cx
39	Brain_MD_39	Medulloblastoma	Classic	T3M0	M	10yr 9m	53	A	V,C,Cx
40	Brain_MD_40	Medulloblastoma	Desmoplastic	T4M3	M	3yr 4m	57	A	V,C,Cx
41	Brain_MD_41	Medulloblastoma	Classic	T4M0	F	4yr 8m	60	A	V,C,Cx,VP
42	Brain_MD_42	Medulloblastoma	Classic	T3M3	M	6yr	62	A	V,C,Cx,VP
43	Brain_MD_43	Medulloblastoma	Classic	T3M0	M	9yr 3m	64	A	V,C,Cx
44	Brain_MD_44	Medulloblastoma	Classic	T3M0	M	5yr 3m	66	A	V,C,Cx
45	Brain_MD_45	Medulloblastoma	Classic	T4M0	M	3yr 6m	68	A	V,C,Cx,P
46	Brain_MD_46	Medulloblastoma	Classic	T3M0	M	2yr 4m	68	A	V,C,Cx
47	Brain_MD_47	Medulloblastoma	Classic	T4M0	F	10yr 6m	70	A	V,C,Cx
48	Brain_MD_48	Medulloblastoma	Classic	T3bM0	M	5yr 5m	72	A	V,C,Cx,VP,Ca
49	Brain_MD_49	Medulloblastoma	Classic	T2M0	F	12yr 11m	74	A	V,C,Cx
50	Brain_MD_50	Medulloblastoma	Classic	T3bM0	M	9yr 11m	79	A	V,C,Cx
51	Brain_MD_51	Medulloblastoma	Classic	T3bM0	M	13yr 8m	79	A	V,C,Cx
52	Brain_MD_52	Medulloblastoma	Classic	T2M0	M	1yr 8m	80	A	V,C,Cx
53	Brain_MD_53	Medulloblastoma	Desmoplastic	T2M0	F	5yr 2m	84	A	V,C,Cx
54	Brain_MD_54	Medulloblastoma	Classic	T4M4	F	1yr 5m	85	A	V,C,Cx,VP,Ca,T,M
55	Brain_MD_55	Medulloblastoma	Classic	T3bM2	M	10yr 4m	87	A	V,C,Cx,VP
56	Brain_MD_56	Medulloblastoma	Desmoplastic	T2M0	F	28yr	87	A	V,C
57	Brain_MD_57	Medulloblastoma	Classic	T2M3	M	2yr 7m	97	A	V,C,Cx
58	Brain_MD_58	Medulloblastoma	Classic	T1M0	M	3yr 7m	108	A	V,C,Cx,VP
59	Brain_MD_59	Medulloblastoma	Classic	T3bM0	M	9yr 9m	130	A	V,C
60	Brain_MD_60	Medulloblastoma	Desmoplastic	T3M0	F	2yr	24	A	V,C,Cx
61	Brain_MD_61	Medulloblastoma							
62	Brain_MD_62	Medulloblastoma							V,C,Cx
63	Brain_MD_63	Medulloblastoma							
64	Brain_MD_64	Medulloblastoma							V,C,Cx
65	Brain_MD_65	Medulloblastoma							V,C,Cx
66	Brain_MD_66	Medulloblastoma							V,C
67	Brain_MD_67	Medulloblastoma							V,C,Cx,VP
68	Brain_MGlio_1	Malignant Glioma							
69	Brain_MGlio_2	Malignant Glioma							V= vincristine
70	Brain_MGlio_3	Malignant Glioma							C= cisplatin
71	Brain_MGlio_4	Malignant Glioma							Cx= cytoxan
72	Brain_MGlio_5	Malignant Glioma							VP= etoposide
73	Brain_MGlio_6	Malignant Glioma							CC= CCNU
74	Brain_MGlio_7	Malignant Glioma							Ca= carboplatin
75	Brain_MGlio_8	Malignant Glioma							P= procarbazine
76	Brain_MGlio_9	Malignant Glioma							M= methotrexate
77	Brain_MGlio_10	Malignant Glioma							T= thiotepa
78	Brain_Rhab_1	AT/RT (Brain)							
79	Brain_Rhab_2	AT/RT (Renal)							
80	Brain_Rhab_3	AT/RT (Renal)							

81	Brain_Rhab_4	AT/RT (Brain)
82	Brain_Rhab_5	AT/RT (Extra Renal)
83	Brain_Rhab_6	AT/RT (Extra Renal)
84	Brain_Rhab_7	AT/RT (Renal)
85	Brain_Rhab_8	AT/RT (Brain)
86	Brain_Rhab_9	AT/RT (Brain)
87	Brain_Rhab_10	AT/RT (Brain)
88	Brain_Ncer_1	Normal cerebellum
89	Brain_Ncer_2	Normal cerebellum
90	Brain_Ncer_3	Normal cerebellum
91	Brain_Ncer_4	cerebellum
92	Brain_PNET_1	PNET
93	Brain_PNET_2	PNET
94	Brain_PNET_3	PNET
95	Brain_PNET_4	PNET
96	Brain_PNET_5	PNET
97	Brain_PNET_6	PNET
98	Brain_PNET_7	PNET (pineoblastoma)
99	Brain_PNET_8	PNET (pineoblastoma)

Dataset A, A1, A2 - multiple tumor samples

Dataset A: 10 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 4 normal cerebellums and 8 supratentorial PNETs.

Two of the supratentorial PNETs are pineoblastomas, which historically have been inconsistently included in the PNET category. The analysis was repeated excluding these 2 pineoblastomas.

Dataset A1: 10 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 4 normal cerebellums and 6 supratentorial PNETs.

To test whether inclusion of a larger number of medulloblastomas might lessen the distinctions noted in Dataset A, 50 more medulloblastoma samples were added and the PCA analysis repeated.

Dataset A2: 60 medulloblastomas, 10 malignant gliomas, 10 AT/RT (5 CNS, 5 renal-extrarenal), 4 normal cerebellums and 6 supratentorial PNETs.

Dataset A

Sample number	Sample name	Type
1	Brain_MD_12	Medulloblastoma
2	Brain_MD_61	Medulloblastoma
3	Brain_MD_15	Medulloblastoma
4	Brain_MD_57	Medulloblastoma
5	Brain_MD_33	Medulloblastoma
6	Brain_MD_64	Medulloblastoma
7	Brain_MD_17	Medulloblastoma
8	Brain_MD_62	Medulloblastoma
9	Brain_MD_63	Medulloblastoma
10	Brain_MD_32	Medulloblastoma
11	Brain_MGlio_1	Malignant Glioma
12	Brain_MGlio_2	Malignant Glioma
13	Brain_MGlio_3	Malignant Glioma
14	Brain_MGlio_4	Malignant Glioma
15	Brain_MGlio_5	Malignant Glioma
16	Brain_MGlio_6	Malignant Glioma
17	Brain_MGlio_7	Malignant Glioma
18	Brain_MGlio_8	Malignant Glioma
19	Brain_MGlio_9	Malignant Glioma
20	Brain_MGlio_10	Malignant Glioma
21	Brain_Rhab_1	AT/RT (Brain)
22	Brain_Rhab_2	AT/RT (Renal)
23	Brain_Rhab_3	AT/RT (Renal)
24	Brain_Rhab_4	AT/RT (Brain)
25	Brain_Rhab_5	AT/RT (Extra Renal)

26	Brain_Rhab_6	AT/RT (Extra Renal)
27	Brain_Rhab_7	AT/RT (Renal)
28	Brain_Rhab_8	AT/RT (Brain)
29	Brain_Rhab_9	AT/RT (Brain)
30	Brain_Rhab_10	AT/RT (Brain)
31	Brain_Ncer_1	Normal cerebellum
32	Brain_Ncer_2	Normal cerebellum
33	Brain_Ncer_3	Normal cerebellum
34	Brain_Ncer_4	Normal cerebellum
35	Brain_PNET_1	PNET
36	Brain_PNET_2	PNET
37	Brain_PNET_3	PNET
38	Brain_PNET_4	PNET
39	Brain_PNET_5	PNET
40	Brain_PNET_6	PNET
41	Brain_PNET_7	PNET (pineoblastoma)
42	Brain_PNET_8	PNET (pineoblastoma)

Dataset A1

Sample number	Sample name	Type
1	Brain_MD_12	Medulloblastoma
2	Brain_MD_61	Medulloblastoma
3	Brain_MD_15	Medulloblastoma
4	Brain_MD_57	Medulloblastoma
5	Brain_MD_33	Medulloblastoma
6	Brain_MD_64	Medulloblastoma
7	Brain_MD_17	Medulloblastoma
8	Brain_MD_62	Medulloblastoma
9	Brain_MD_63	Medulloblastoma

10	Brain_MD_32	Medulloblastoma
11	Brain_MGlio_1	Malignant Glioma
12	Brain_MGlio_2	Malignant Glioma
13	Brain_MGlio_3	Malignant Glioma
14	Brain_MGlio_4	Malignant Glioma
15	Brain_MGlio_5	Malignant Glioma
16	Brain_MGlio_6	Malignant Glioma
17	Brain_MGlio_7	Malignant Glioma
18	Brain_MGlio_8	Malignant Glioma
19	Brain_MGlio_9	Malignant Glioma
20	Brain_MGlio_10	Malignant Glioma
21	Brain_Rhab_1	AT/RT (Brain)
22	Brain_Rhab_2	AT/RT (Renal)
23	Brain_Rhab_3	AT/RT (Renal)
24	Brain_Rhab_4	AT/RT (Brain)
25	Brain_Rhab_5	AT/RT (Extra Renal)
26	Brain_Rhab_6	AT/RT (Extra Renal)
27	Brain_Rhab_7	AT/RT (Renal)
28	Brain_Rhab_8	AT/RT (Brain)
29	Brain_Rhab_9	AT/RT (Brain)
30	Brain_Rhab_10	AT/RT (Brain)
31	Brain_Ncer_1	Normal cerebellum
32	Brain_Ncer_2	Normal cerebellum
33	Brain_Ncer_3	Normal cerebellum
34	Brain_Ncer_4	Normal cerebellum
35	Brain_PNET_1	PNET
36	Brain_PNET_2	PNET
37	Brain_PNET_3	PNET
38	Brain_PNET_4	PNET
39	Brain_PNET_5	PNET

40	Brain_PNET_6	PNET

Dataset A2

Sample number	Sample name	Type
1	Brain_MD_1	Medulloblastoma
2	Brain_MD_2	Medulloblastoma
3	Brain_MD_3	Medulloblastoma
4	Brain_MD_4	Medulloblastoma
5	Brain_MD_5	Medulloblastoma
6	Brain_MD_6	Medulloblastoma
7	Brain_MD_7	Medulloblastoma
8	Brain_MD_8	Medulloblastoma
9	Brain_MD_9	Medulloblastoma
10	Brain_MD_10	Medulloblastoma
11	Brain_MD_11	Medulloblastoma
12	Brain_MD_12	Medulloblastoma
13	Brain_MD_13	Medulloblastoma
14	Brain_MD_14	Medulloblastoma
15	Brain_MD_15	Medulloblastoma
16	Brain_MD_16	Medulloblastoma
17	Brain_MD_17	Medulloblastoma
18	Brain_MD_18	Medulloblastoma
19	Brain_MD_19	Medulloblastoma
20	Brain_MD_20	Medulloblastoma
21	Brain_MD_21	Medulloblastoma
22	Brain_MD_22	Medulloblastoma
23	Brain_MD_23	Medulloblastoma
24	Brain_MD_24	Medulloblastoma
25	Brain_MD_25	Medulloblastoma
26	Brain_MD_26	Medulloblastoma

27	Brain_MD_27	Medulloblastoma
28	Brain_MD_28	Medulloblastoma
29	Brain_MD_29	Medulloblastoma
30	Brain_MD_30	Medulloblastoma
31	Brain_MD_31	Medulloblastoma
32	Brain_MD_32	Medulloblastoma
33	Brain_MD_33	Medulloblastoma
34	Brain_MD_34	Medulloblastoma
35	Brain_MD_35	Medulloblastoma
36	Brain_MD_36	Medulloblastoma
37	Brain_MD_37	Medulloblastoma
38	Brain_MD_38	Medulloblastoma
39	Brain_MD_39	Medulloblastoma
40	Brain_MD_40	Medulloblastoma
41	Brain_MD_41	Medulloblastoma
42	Brain_MD_42	Medulloblastoma
43	Brain_MD_43	Medulloblastoma
44	Brain_MD_44	Medulloblastoma
45	Brain_MD_45	Medulloblastoma
46	Brain_MD_46	Medulloblastoma
47	Brain_MD_47	Medulloblastoma
48	Brain_MD_48	Medulloblastoma
49	Brain_MD_49	Medulloblastoma
50	Brain_MD_50	Medulloblastoma
51	Brain_MD_51	Medulloblastoma
52	Brain_MD_52	Medulloblastoma
53	Brain_MD_53	Medulloblastoma
54	Brain_MD_54	Medulloblastoma
55	Brain_MD_55	Medulloblastoma
56	Brain_MD_56	Medulloblastoma

57	Brain_MD_57	Medulloblastoma
58	Brain_MD_58	Medulloblastoma
59	Brain_MD_59	Medulloblastoma
60	Brain_MD_60	Medulloblastoma
61	Brain_MGlio_1	Malignant Glioma
62	Brain_MGlio_2	Malignant Glioma
63	Brain_MGlio_3	Malignant Glioma
64	Brain_MGlio_4	Malignant Glioma
65	Brain_MGlio_5	Malignant Glioma
66	Brain_MGlio_6	Malignant Glioma
67	Brain_MGlio_7	Malignant Glioma
68	Brain_MGlio_8	Malignant Glioma
69	Brain_MGlio_9	Malignant Glioma
70	Brain_MGlio_10	Malignant Glioma
71	Brain_Rhab_1	AT/RT (Brain)
72	Brain_Rhab_2	AT/RT (Renal)
73	Brain_Rhab_3	AT/RT (Renal)
74	Brain_Rhab_4	AT/RT (Brain)
75	Brain_Rhab_5	AT/RT (Extra Renal)
76	Brain_Rhab_6	AT/RT (Extra Renal)
77	Brain_Rhab_7	AT/RT (Renal)
78	Brain_Rhab_8	AT/RT (Brain)
79	Brain_Rhab_9	AT/RT (Brain)
80	Brain_Rhab_10	AT/RT (Brain)
81	Brain_Ncer_1	Normal cerebellum
82	Brain_Ncer_2	Normal cerebellum
83	Brain_Ncer_3	Normal cerebellum
84	Brain_Ncer_4	Normal cerebellum
85	Brain_PNET_1	PNET
86	Brain_PNET_2	PNET

87	Brain_PNET_3	PNET
88	Brain_PNET_4	PNET
89	Brain_PNET_5	PNET
90	Brain_PNET_6	PNET

Dataset B - MD classic-desmoplastic

Dataset B: 25 classic and 9 desmoplastic medulloblastomas.

Number	Sample name	Type	Subtype
1	Brain_MD_7	Medulloblastoma	Classic
2	Brain_MD_59	Medulloblastoma	Classic
3	Brain_MD_20	Medulloblastoma	Classic
4	Brain_MD_21	Medulloblastoma	Classic
5	Brain_MD_50	Medulloblastoma	Classic
6	Brain_MD_49	Medulloblastoma	Classic
7	Brain_MD_45	Medulloblastoma	Classic
8	Brain_MD_43	Medulloblastoma	Classic
9	Brain_MD_8	Medulloblastoma	Classic
10	Brain_MD_42	Medulloblastoma	Classic
11	Brain_MD_1	Medulloblastoma	Classic
12	Brain_MD_4	Medulloblastoma	Classic
13	Brain_MD_55	Medulloblastoma	Classic
14	Brain_MD_41	Medulloblastoma	Classic
15	Brain_MD_37	Medulloblastoma	Classic
16	Brain_MD_3	Medulloblastoma	Classic
17	Brain_MD_34	Medulloblastoma	Classic
18	Brain_MD_29	Medulloblastoma	Classic
19	Brain_MD_13	Medulloblastoma	Classic
20	Brain_MD_24	Medulloblastoma	Classic

21	Brain_MD_65	Medulloblastoma	Classic
22	Brain_MD_5	Medulloblastoma	Classic
23	Brain_MD_66	Medulloblastoma	Classic
24	Brain_MD_67	Medulloblastoma	Classic
25	Brain_MD_58	Medulloblastoma	Classic
26	Brain_MD_53	Medulloblastoma	Desmoplastic
27	Brain_MD_56	Medulloblastoma	Desmoplastic
28	Brain_MD_16	Medulloblastoma	Desmoplastic
29	Brain_MD_40	Medulloblastoma	Desmoplastic
30	Brain_MD_35	Medulloblastoma	Desmoplastic
31	Brain_MD_30	Medulloblastoma	Desmoplastic
32	Brain_MD_23	Medulloblastoma	Desmoplastic
33	Brain_MD_28	Medulloblastoma	Desmoplastic
34	Brain_MD_60	Medulloblastoma	Desmoplastic

Dataset C - MD outcome

Dataset C: 39 medulloblastomas survivors and 21 treatment failures (non-survivors)

Number	Sample name	Type	Subtype	Chang Stage	Sex	Age at diagnosis [years/ months]	Followup [Months]	Current status [Alive/Dead]	Chemotherapy
1	Brain_MD_1	Medulloblastoma	Classic	T4M1	M	8m	11	D	V,C,Cx,VP
2	Brain_MD_2	Medulloblastoma	Classic	T2M0	M	8yr10m	5	D	V,C,Cx,VP
3	Brain_MD_3	Medulloblastoma	Classic	T3M0	M	6yr	7	D	V,C,Cx
4	Brain_MD_4	Medulloblastoma	Classic	T3M3	M	5yr 3m	7	D	V,C,Cx,VP
5	Brain_MD_5	Medulloblastoma	Classic	M3	M	38yr 2m	7	D	V,C
6	Brain_MD_6	Medulloblastoma	Classic	T4M0	F	7m	9	D	V,C,Cx
7	Brain_MD_7	Medulloblastoma	Classic	T1M0	M	6yr 5m	14	D	V,C,Cx
8	Brain_MD_8	Medulloblastoma	Classic	T3bM1	M	6yr 1m	16	D	V,C,Cx
9	Brain_MD_9	Medulloblastoma	Classic	M0	M	8yr	18	D	V,C,Cx,VP
10	Brain_MD_10	Medulloblastoma	Classic	M0	M	3yr 10m	18	D	V,C,Cx
11	Brain_MD_11	Medulloblastoma	Classic	T2M1	M	8yr 2m	19	D	V,C,Cx,VP,Ca,T,M
12	Brain_MD_12	Medulloblastoma	Classic	M0	F	3yr 9m	25	D	V,C,Cx
13	Brain_MD_13	Medulloblastoma	Classic	T3M3	M	14yr 5m	26	D	V,C,Cx
14	Brain_MD_14	Medulloblastoma	Desmoplastic	M0	M	6yr 3m	33	D	V,C,CC
15	Brain_MD_15	Medulloblastoma	Desmoplastic	T2MO	F	11yr 7m	38	D	V,C,Cx,VP
16	Brain_MD_16	Medulloblastoma	Desmoplastic	T3M3	F	11yr 5m	39	D	V,C,VP
17	Brain_MD_17	Medulloblastoma	Classic	T3bM3	F	3yr 3m	39	D	V,C,Cx
18	Brain_MD_18	Medulloblastoma	Classic	T2M3	M	4yr 4m	42	D	V,C,Cx
19	Brain_MD_19	Medulloblastoma	Classic	M2	F	26yr 1m	65	D	V,C,Cx,VP
20	Brain_MD_20	Medulloblastoma	Classic	T3bM0	M	20yr 6m	92	D	V,C

21	Brain_MD_21	Medulloblastoma	Classic	T2M0	F	23yr 3m	102	D	V,C
22	Brain_MD_22	Medulloblastoma	Desmoplastic	M0	F	5yr 7m	24	A	V,C,CC
23	Brain_MD_23	Medulloblastoma	Desmoplastic	T4M0	M	1yr 4m	25	A	V,C,Cx
24	Brain_MD_24	Medulloblastoma	Classic	T3M0	M	10yr 10m	27	A	V,C,Cx
25	Brain_MD_25	Medulloblastoma	Classic	M0	F	5yr 4m	28	A	V,C,Cx,VP
26	Brain_MD_26	Medulloblastoma	Classic	T2M3	M	1yr	33	A	V,C,Cx,VP
27	Brain_MD_27	Medulloblastoma	Classic	M0	M	5yr 10m	34	A	V,C,Cx
28	Brain_MD_28	Medulloblastoma	Desmoplastic	T4M0	M	6yr 1m	35	A	V,C,Cx
29	Brain_MD_29	Medulloblastoma	Classic	T3M0	F	7yr 5m	35	A	V,C,Cx
30	Brain_MD_30	Medulloblastoma	Desmoplastic	T3M0	F	11yr 9m	36	A	V,C,Cx
31	Brain_MD_31	Medulloblastoma	Classic	M0	M	7yr 4m	39	A	V,C,Cx
32	Brain_MD_32	Medulloblastoma	Desmoplastic	T2M0	M	10yr 11m	39	A	V,C,Cx
33	Brain_MD_33	Medulloblastoma	Classic	T3bM0	M	12yr 9m	41	A	V,C,Cx
34	Brain_MD_34	Medulloblastoma	Classic	T3M1	M	8yr 2m	42	A	V,C,Cx
35	Brain_MD_35	Medulloblastoma	Desmoplastic	T3M0	F	2yr 3m	45	A	V,C,Cx
36	Brain_MD_36	Medulloblastoma	Classic	T3M0	M	5yr 6m	46	A	V,C,Cx
37	Brain_MD_37	Medulloblastoma	Classic	T3M0	F	12yr 7m	51	A	V,C,Cx
38	Brain_MD_38	Medulloblastoma	Desmoplastic	T3M1	F	7m	52	A	V,C,Cx
39	Brain_MD_39	Medulloblastoma	Classic	T3M0	M	10yr 9m	53	A	V,C,Cx
40	Brain_MD_40	Medulloblastoma	Desmoplastic	T4M3	M	3yr 4m	57	A	V,C,Cx
41	Brain_MD_41	Medulloblastoma	Classic	T4M0	F	4yr 8m	60	A	V,C,Cx,VP
42	Brain_MD_42	Medulloblastoma	Classic	T3M3	M	6yr	62	A	V,C,Cx,VP
43	Brain_MD_43	Medulloblastoma	Classic	T3M0	M	9yr 3m	64	A	V,C,Cx
44	Brain_MD_44	Medulloblastoma	Classic	T3M0	M	5yr 3m	66	A	V,C,Cx
45	Brain_MD_45	Medulloblastoma	Classic	T4M0	M	3yr 6m	68	A	V,C,Cx,P
46	Brain_MD_46	Medulloblastoma	Classic	T3M0	M	2yr 4m	68	A	V,C,Cx
47	Brain_MD_47	Medulloblastoma	Classic	T4M0	F	10yr 6m	70	A	V,C,Cx
48	Brain_MD_48	Medulloblastoma	Classic	T3bM0	M	5yr 5m	72	A	V,C,Cx,VP,Ca
49	Brain_MD_49	Medulloblastoma	Classic	T2M0	F	12yr 11m	74	A	V,C,Cx
50	Brain_MD_50	Medulloblastoma	Classic	T3bM0	M	9yr 11m	79	A	V,C,Cx
51	Brain_MD_51	Medulloblastoma	Classic	T3bM0	M	13yr 8m	79	A	V,C,Cx
52	Brain_MD_52	Medulloblastoma	Classic	T2M0	M	1yr 8m	80	A	V,C,Cx
53	Brain_MD_53	Medulloblastoma	Desmoplastic	T2M0	F	5yr 2m	84	A	V,C,Cx
54	Brain_MD_54	Medulloblastoma	Classic	T4M4	F	1yr 5m	85	A	V,C,Cx,VP,Ca,T,M
55	Brain_MD_55	Medulloblastoma	Classic	T3bM2	M	10yr 4m	87	A	V,C,Cx,VP
56	Brain_MD_56	Medulloblastoma	Desmoplastic	T2M0	F	28yr	87	A	V,C
57	Brain_MD_57	Medulloblastoma	Classic	T2M3	M	2yr 7m	97	A	V,C,Cx
58	Brain_MD_58	Medulloblastoma	Classic	T1M0	M	3yr 7m	108	A	V,C,Cx,VP
59	Brain_MD_59	Medulloblastoma	Classic	T3bM0	M	9yr 9m	130	A	V,C
60	Brain_MD_60	Medulloblastoma	Desmoplastic	T3M0	F	2yr	24	A	V,C,Cx

Please replace the paragraph beginning at page 5, line 23 through page 6, line 14 with the following amended paragraph:

Figs. 1A-1I are depictions of methods and data obtained in classifying embryonal brain tumors by gene expression. Fig. 1A-1E show representative photomicrographs of embryonal and non-embryonal tumors: 1A) classic medulloblastoma, 1B) desmoplastic medulloblastoma, 1C) supratentorial primitive neuroectodermal tumor (PNET), 1D) atypical teratoid/rhabdoid tumor

(AT/RT; arrow indicates rhabdoid cell morphology), and 1E) glioblastoma with pseudopalisading necrosis (n). Fig. 1F is a schematic representation of principal component analysis (PCA) of tumor samples using all genes exhibiting variation across the dataset. The axes represent the 3 linear combinations of genes that account for the majority of the variance in the original dataset ~~(see Supplementary Information Section I and III at the world wide web site:; genome.wi.mit.edu/MPR/CNS)~~. Fig. 1G is a schematic representation of PCA using 50 genes selected by signal-to-noise metric to be most highly associated each tumor type (the top 10 for each tumor are listed in Fig. 1I). Fig. 1H is a schematic representation of clustering of tumor samples by hierarchical clustering using all genes exhibiting variation across the dataset. Fig. 1I is a graphical representation of signal-to-noise rankings of genes comparing each tumor type to all other types combined ~~(see Supplementary Information Section I; http://www.genome.wi.mit.edu/MPR/CNS])~~. For each gene, red indicates high level of expression relative to the mean, blue indicates low level of expression relative to the mean. The gene names for Fig. 1I are shown in Table 4.

Please replace the paragraph at page 6, line 15 through 23 with the following amended paragraph:

Figs. 2A-2C are graphical representations of differential expression of genes in classic versus desmoplastic medulloblastomas. Depicted are data used to rank Genes by the signal-to-noise metric according to their correlation with the classic vs. desmoplastic distinction. Genes shown are those more highly correlated with the distinction than 99% of permutations of the class labels ($p < 0.01$~~; see Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS; the entire teachings of which are incorporated herein by reference~~). GenBank accession numbers and gene descriptions are shown. Genes regulated by *Shh* are shown at right. The gene names for Figs. 2A-2B are shown in Table 5.

Please replace the paragraph at page 37, lines 3 through 16 with the follwing amended paragraph:

The problem of distinguishing different embryonal CNS tumors from each other was addressed. This is important because the classification of these tumors based on histopathological appearance is debated (Figs. 1A-1E). Some argue that medulloblastomas are part of a larger class of PNETs arising from a common cell type in the subventricular germinal matrix, whereas others believe that they arise from cerebellar granule cell progenitors (Rorke, L., 1983. *J. Neuropathol. Exp. Neurol.*, 42:1-15; Kadin, M. *et al.*, 1970. *J. Neuropath. Exp. Neurol.*, 29:583-600). To begin to generate a molecular taxonomy of CNS embryonal tumors, the gene expression profiles of 42 patient samples were analyzed (Set A: 10 medulloblastomas, 5 CNS AT/RT, 5 renal and extrarenal rhabdoid tumors, and 8 supratentorial PNETs, as well as 10 non-embryonal brain tumors (malignant glioma) and 4 normal human cerebella). RNA extracted from frozen specimens was analyzed with oligonucleotide microarrays containing probes for 6817 genes. The gene expression data are available in "Section II" below ~~of "Supplementary Information" (http://www.genome.wi.mit.edu/MPR/CNS)~~.

Please replace the paragraph at page 37, line 17 through page 37, line 7 with the following amended paragraph:

To determine whether the different types of tumors could be molecularly distinguished, a method of data reduction known as "Principal Component Analysis" in which the high dimensionality of the data was reduced to 3 viewable dimensions representing linear combinations of variables (genes) that account for the majority of the variance in the original dataset was used (Fig. 1F; Mardia, K. *et al.*, 1979. Multivariate Analysis. Academic Press London.). Normal brain was easily separable from the brain tumors and the different tumor types were similarly separable. Separation of tumor types was also seen using hierarchical clustering (Fig. 1H; Eisen, M. *et al.*, 1998. *Proc. Natl. Acad. Sci. USA*, 95:14863-14868). A more appropriate strategy for distinguishing known tumor types, however, is to use supervised learning methods to identify the genes most highly correlated with the tumor type distinctions (Figs. 1G and 1I, and Table 4). Analysis of 1,000 random permutations of the data failed to yield a separation of tumor classes to the extent observed in Fig. 1G, indicating that the observed gene

expression patterns could not be explained by chance ~~(Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS)~~. The robustness of these markers for classification was further investigated using a Weighted Voting algorithm and evaluated by cross validation testing (Golub, T. *et al.*, 1999. *Science*, 286:531-537). Correct classification of the tumors was achieved with accuracy (35 of 42 correct classifications, $P < 10^{-10}$ compared to random classification~~; Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS~~).

Please replace the paragraph at page 38, lines 8 through 23 with the following amended paragraph:

As expected, malignant gliomas were clearly separable from medulloblastomas, reflecting the derivation of gliomas from cells of non-neuronal origin. Consistent with this, the gliomas expressed genes typical of the astrocytic and oligodendrocytic lineage (*PEA-15*, *SOX2*, *PMP-2*, *Olig-2*, *TrkB* kinase-negative splice variant, *S-100*, *GFAP*), genes related to metabolism (fructose 2,6-bisphosphatase, glutamate dehydrogenase), and genes involved in cell differentiation (*ID2*, *GDF-1*, *TYK2*; Fig. 1I and Table 4~~, and Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS~~). Unexpectedly, the medulloblastomas form a cluster that is also separate from the PNETs (Fig. 1G), supporting the notion that these two classes of embryonal tumors are indeed molecularly distinct. Among the genes most highly correlated with the medulloblastoma class were *Zic* and *NSCL-1*, encoding transcription factors that have been shown to be specific for cerebellar granule cells (Fig. 1I and Table 4; Aruga, J. *et al.*, 1994. *J. Neurochem.*, 63:1880-1890; Yokota, N. *et al.*, 1996. Cancer Res., 56:377-383). This result suggests that medulloblastomas, but not PNETs, arise from cerebellar granule cells, or alternatively, have activated the transcriptional program of cerebellar granule cells.

Please replace the paragraph at page 38, line 24 through page 39, line 11 with the following amended paragraph:

Accurate identification of AT/RT is also important because patients with these tumors have an extremely poor prognosis. AT/RT arise either in the CNS or in other organs such as the kidney, where they are referred to as rhabdoid tumors. Most tumors harbor hSNF5/INI1 mutations, but it is unknown whether AT/RT arising in different anatomical locations are molecularly distinct (Rorke, L. *et al.*, 1996. *J. Neurosurg.*, 85:56-65; Biegel, J. *et al.*, 1999. *Cancer Res.*, 59:74-79; Versteege, I. *et al.*, 1998. *Nature*, 394:203-6). As shown in Fig. 1G, the AT/RT and rhabdoid tumors were clearly distinguishable from the other tumor types in the study. Strikingly, the CNS AT/RT and abdominal rhabdoid tumors were molecularly similar despite having arisen in different anatomical locations. This finding supports the notion that they arise from a similar cell of origin. Alternatively, a common mechanism of transformation yield similar transcriptional programs in cells of distinct origin. Markers of the AT/RT/rhabdoid distinction include genes specifically expressed during myogenesis, including skeletal β-tropomyosin, neutral calponin, *NF-AT3*, myosin regulatory light chain (Fig. 1I and Table 4~~, and Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS~~). This finding is consistent with the notion that the tumors have a mesenchymal origin.

Please replace the paragraph at page 39, line 25 through page 40, line 15 with the following amended paragraph:

To determine whether desmoplastic and classic medulloblastoma are distinguishable by gene expression, 34 medulloblastoma samples (Set B) whose histology was scored using World Health Organization criteria were analyzed (Giangaspero, F. *et al.*, 2000. Medulloblastoma. In: Kleihues, P. and Cavenee, W. (eds.). World Health Organization Histological Classification of Tumours of the Nervous System. Lyon: International Agency for Research on Cancer, pp. 129-137). As shown in Table 5 and Figs. 2A and 2B, a sharp and statistically significant gene expression signature of desmoplastic histology was evident, and this signature was sufficient for correct classification of 33 of 34 tumors ($P = 8.6 \times 10^{-7}$ compared to random classification~~; Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS~~). Strikingly, among the genes most highly correlated with desmoplastic medulloblastoma, see Fig. 2C, were *PTCH* (itself a transcriptional target of Shh) as well as two other *Shh* downstream targets: *Gli*

and *N-Myc* (Murone, M. *et al.*, 1999. *Curr. Biol.*, 28:76-84). Furthermore, *IGF2* expression was correlated with desmoplastic histology, and its expression is known to be essential for Shh-mediated tumorigenesis in mice (Hahn, H. *et al.*, 2000. *J. Biol. Chem.*, 275:28341-28344). Taken together, the transcriptional profiling indicates that sporadic desmoplastic medulloblastomas, like Gorlin's syndrome-associated tumors, are characterized by activation of Shh signaling pathway, further supporting the suspicion that Shh dysregulation may be important in the pathogenesis of medulloblastoma.

Please replace the paragraph at page 40, line 28 through page 41, line 21 with the following amended paragraph:

To explore the heterogeneity in medulloblastoma treatment response, the analysis was expanded to include 60 similarly treated patients from whom biopsies were obtained prior to receiving treatment, and for whom clinical follow-up was available (Set C). Clustering methods were first used to determine if they would identify biologically distinct subsets of the tumors. The tumors were clustered into two groups using Self-Organizing Maps (SOMs), an unsupervised algorithm that groups samples into a predetermined number of clusters based on their gene expression patterns (Golub, T. *et al.*, 1999. *Science*, 286:531-537; Tamayo, P. *et al.*, 1999. *Proc. Natl. Acad. Sci. USA*, 96:2907-2912). The genes most highly correlated with the SOM clusters were primarily ribosomal protein-encoding genes ~~(Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS)~~, suggesting differences in ribosome biogenesis. Blinded electron microscopic examination of 9 samples by 3 observers confirmed that tumors falling into the cluster characterized by high expression of ribosomal protein genes indeed contained higher numbers of ribosomes (P = 0.03, Fisher exact test). The next question was whether the SOM-derived clusters were correlated with patient survival. No statistically significant difference in the proportion of survivors versus treatment failures in each cluster was observed (Fisher Exact Test P = 0.1~~; Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS~~). A supervised learning gene expression-based outcome predictor was developed in which the classifier 'learns' the distinction between patients who are alive following treatment ('survivors') compared to those who succumbed to their

disease ('failures'; minimum follow-up 24 months for surviving patients; overall median 41.5 months).

Please replace the paragraph at page 42, lines 3 through 25 with the following amended paragraph:

Gene expression-based outcome predictions were statistically significant for k-NN models ranging from 2 to 21 genes, with optimal predictions made by an 8-gene model which made only 13/60 classification errors (Fisher Exact Test P = 0.0002). Shown most clearly by Kaplan-Meier survival analysis in Fig. 3A, patients predicted to be Survivors had a 5-year overall survival of 80% compared to 17% for patients predicted to have a poor outcome (P = 0.000003, log-rank test). A more conservative method of assessing statistical significance is to attempt to optimize classifiers of random permutations of the Survivor/Failure class labels. 1000 such permutations were determined, and only 9/1000 permutations were found for which prediction accuracy matched or exceeded our observed result ~~(Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS)~~, indicating that the result is unlikely to be achieved by chance (P = 0.009). Therefore, several other classification algorithms including Weighted Voting were subsequently tested (Golub, T. *et al.*, 1999. *Science*, 286:531-537; Slonim, D. *et al.*, 2000. *Procs. of the Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, Japan April 8 - 11, p263-272, 2000), Support Vector Machines (Mukherjee, S. *et al.*, 1999. Support vector machine classification of microarray data. CBCL Paper #182/AI Memo #1676, Massachusetts Institute of Technology, Cambridge, MA; Brown, M. *et al.*, 2000. *Proc. Natl. Acad. Sci. USA*, 97:262-267), and IBM SPLASH (Califano *et al.*, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, San Diego, California, August 19-23, p75-85, 1999), all of which performed with similarly high accuracy ~~(Supplementary Information, Sections I and III; http://www.genome.wi.mit.edu/MPR/CNS)~~.

Please replace the paragraph at page 42, line 26 through page 43, line 14 with the following amended paragraph:

The clinical value of the predictor was explored further by considering existing prognostic factors for medulloblastoma outcome. Patients with localized disease (M0) had a more favorable outcome compared to patients with involvement of the cerebrospinal fluid or with distant metastases (M+) (P = 0.03 comparing M0 with M+ by Kaplan-Meier analysis), although not all M0 patients survived. When the outcome predictor was applied only to the 42 M0 patients, the prediction of outcome remained significant (P = 0.002), indicating that the expression-based predictor substantially improved staging-based prognostication. Similarly, *TrkC*-based prediction was imperfect in this series in that not all patients in the unfavorable (*TrkC*-low) category died. When the gene expression-based predictor was applied to the 33 *TrkC*-low patients, the surviving patients could be significantly separated from those who succumbed to their disease (P = 0.01~~; Supplementary Information Section III; http://www.genome.wi.mit.edu/MPR/CNS~~). Of note, not all patients in this study received identical therapy. However, restricting the analysis to the 35 patients that received surgery, vincristine, cisplatin and cyclophosphamide, the predictor continued to yield a significant Kaplan-Meier survival distinction (P = 0.0012). Taken together, these results demonstrate that the gene expression-based outcome predictor exceeds other approaches to prognosis determination.

Please replace the paragraph at page 44, line 7 through page 45, line 8 with the following amended paragraph:

Patient Samples. Patients included 60 children with medulloblastoma, 10 young adults with malignant glioma (WHO grades III and IV), 5 children with AT/RT, 5 with renal/extrarenal rhabdoid tumors, and 8 children with supratentorial PNET (see "expanded methods" below ~~Supplementary Information Section I; http://www.genome.wi.mit.edu/MPR/CNS~~). Medulloblastoma patients were treated with craniospinal irradiation to 2400 - 3600 centiGray (cGy) with a tumor dose of 5300 - 7200 cGy. All patients with medulloblastoma were treated with chemotherapy consisting of cisplatin and vincristine, plus combinations of carboplatin, etoposide, cyclophosphamide or lumustine (CCNU) ~~(details in Supplementary Information Section II; http://www.genome.wi.mit.edu/MPR/CNS)~~. Samples were snap frozen in liquid

nitrogen and stored at -80°C. Studies were done with approval of the Committee for Clinical Investigation of Boston Children's Hospital. The data were organized into three sets: Dataset A (42 samples containing 10 medulloblastoma, 10 malignant glioma, 10 AT/RT, 8 PNET and 4 normal cerebellum), Dataset B (34 samples, containing 9 desmoplastic medulloblastoma and 25 classic medulloblastoma), and Dataset C (60 samples, containing 39 medulloblastoma survivors and 21 treatment failures). The clinical attributes of each of the patients in the study are described ~~available~~ in ~~Supplementary Information~~ Section II below ~~(http://www.genome.wi.mit.edu/MPR/CNS)~~. Tissues were homogenized in guanidinium isothiocyanate and RNA was isolated by centrifugation over a CsCl gradient. RNA integrity was assessed either by northern blotting or by gel electrophoresis. 10-12 μg total RNA was used to generate biotinlylated antisense RNAs which were hybridized overnight to HuGeneFL arrays containing 5920 known genes and 897 expressed sequence tags as previously described (Golub, T. *et al.*, 1999. *Science*, 286:531-537). Arrays were scanned on Affymetrix scanners and the expression value for each gene was calculated using Affymetrix GENECHIP software. Minor differences in microarray intensity were corrected using a linear scaling method as detailed in "expanded methods" below ~~Supplementary Information Section I (http://www.genome.wi.mit.edu/MPR/CNS)~~. Scans were rejected if the scaling factor exceeded 3, fewer than 1000 genes received 'Present' calls, or microarray artifacts were visible.

C12

Please replace the paragraph at page 45, lines 9 through 12 with the following amended paragraph:

Data Analysis: Preprocessing. The gene expression data were subjected to a variation filter that excluded genes showing minimal variation across the samples being analyzed, as detailed in "expanded methods" below ~~Supplementary Information Section I (http://www.genome.wi.mit.edu/MPR/CNS)~~.

C13

Please replace the paragraph at page 45, line 20 through page 46, line 18 with the following amended paragraph:

Data Analysis: Supervised Learning. Genes correlated with particular class distinctions (*e.g.*, classic vs. desmoplastic medulloblastoma) were identified by sorting all of the genes on the array according the signal-to-noise statistic $(\mu_0 - \mu_1)/(\sigma_0 + \sigma_1)$, where μ and σ represent the median and standard deviation of expression, respectively, for each class. Similar results were obtained using a standard t-statistic as the metric $((\mu_0 - \mu_1)/\text{sqrt}(\sigma_0^2/N0 + \sigma_1^2/N_1))$, where N represents the number of samples in each class ~~(see Supplementary Information; http://www.genome.wi.mit.edu/MPR/CNS)~~. Permutation of the column (sample) labels was performed to compare these correlations to what would be expected by chance in 99% of the permutations. For classification, a modification of the k-NN algorithm was developed that predicts the class of a new data point by calculating the Euclidean distance (d) of the new sample to the k nearest samples (for these experiments, k = 5) in the training set using normalized gene expression data, and selecting the class to be that of the majority of the k samples. The weight given to each neighbor was 1/d. The k-NN models were evaluated by 60-fold leave-one-out cross-validation whereby a training set of 59 samples was used to predict the class of a randomly withheld sample, and the cumulative error rate was recorded. Models with variable numbers of genes (1-200, selected according to their correlation with the survivor vs. treatment failure distinction in the training set) were tested in this manner. An 8-gene k-NN outcome prediction model yielded the lowest error rate, and was therefore used to generate Kaplan-Meier survival plots using S-Plus. Predictors using metastatic staging or *TrkC* were constructed by finding the decision boundary half way between the classes: $(\mu_{class0} + \mu_{class1})/2$ using either the staging values 0 vs. 1, 2, 3, 4 or the continuous *TrkC* microarray gene expression levels, and then predicting the unknown sample according to its location with respect to that boundary.